Visual interestingness in image sequences

Grabner, Helmut ; Nater, Fabian ; Druey, Michel D ; Van Gool, Luc

Abstract: Interestingness is said to be the power of attracting or holding one's attention (because something is unusual or exciting, etc.). We, as humans, have the great capacity to direct our visual attention and judge the interestingness of a scene. Consider for example the image sequence in the figure on the right. The spider in front of the camera or the snow on the lens are examples of events that deviate from the context since they violate the expectations, and therefore are considered interesting. On the other hand, weather changes or a camera shift, do not raise human attention considerably, even though large regions of the image are influenced. In this work we firstly investigate what humans consider as "interesting" in image sequences. Secondly we propose a computer vision algorithm to automatically spot these interesting events. To this end, we integrate multiple cues inspired by cognitive concepts and discuss why and to what extent the automatic discovery of visual interestingness is possible.

DOI: https://doi.org/10.1145/2502081.2502109

Visual Interestingness in Image Sequences

Helmut Grabner[1,2], Fabian Nater[1,2], Michel Druey[3], Luc Van Gool[1,4]

[1]Computer Vision Lab, ETH Zurich
{grabner, nater, vangool}@vision.ee.ethz.ch

[2]upicto GmbH
{grabner, nater}@upicto.com

[3]Cognitive Psychology Unit, University of Zurich
m.druey@psychologie.uzh.ch

[4]ESAT - PSI / IBBT, K.U. Leuven
luc.vangool@esat.kuleuven.be

ABSTRACT

Interestingness is said to be the power of attracting or holding one's attention (because something is unusual or exciting, etc.).[1] We, as humans, have the great capacity to direct our visual attention and judge the interestingness of a scene. Consider for example the image sequence in the figure on the right. The spider in front of the camera or the snow on the lens are examples of events that deviate from the context since they violate the expectations, and therefore are considered interesting. On the other hand, weather changes or a camera shift, do not raise human attention considerably, even though large regions of the image are influenced. In this work we firstly investigate what humans consider as "interesting" in image sequences. Secondly we propose a computer vision algorithm to automatically spot these interesting events. To this end, we integrate multiple cues inspired by cognitive concepts and discuss why and to what extent the automatic discovery of visual interestingness is possible.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

computer vision; image understanding; visual interestingness

[1]`http://www.thefreedictionary.com/Interestingness`, 2012/11/08.

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10
http://dx.doi.org/10.1145/2502081.2502109 ...$15.00.

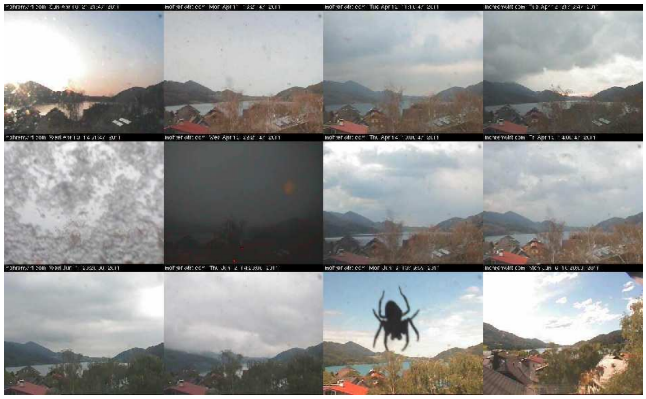


Figure 1: What makes an specific moment in an image sequence interesting and how can we computationally approach this question?

1. INTRODUCTION

What do we mean if we find something "interesting"? This frequently used expression is referred to in very broad, often highly subjective terms. Let us consider three examples to illustrate the concept. The scene in which the airplanes crash into the twin towers during the terrorist attack at 9/11, 2001 shockingly shows that such highly unexpected events strongly drawn our attention. At the moment when it happened, many of us were fixed to the TV-screens, with a mixture of disbelief, disgust and sympathy for the people in the towers. But up to this day, the images evoke great interest, even after seeing them many times and now knowing exactly what will happen. In contrast, the last minutes of a super-bowl final are extremely interesting and capture the attention of many people. This event gains its attraction from the fact that we don't know how the game will end; but loses much of its relevance to most as soon as the game is over. Finally, human interest is raised from very personal experiences. For example watching one's own child playing soccer is much more interesting for the parents than for outsiders.

These examples illustrate that interestingness highly depends on the context, but also on personal experiences and

preferences, which makes it challenging to approach the concept in a principled manner (*cf.*, [23]). Nonetheless, there are scenes that raise widespread interest, whereas others are boring to a majority of people. Humans have a tremendous capacity to assess how interesting a scene or event is and this greatly helps us to navigate through our daily lives. In order to learn more about human visual perception, but also for commercial purposes (*e.g.*, the placement of advertisements), it is of great concern to understand what triggers human attention and interest.

In this work, we restrict ourselves to a particular category of visual input – image sequences recorded by a static video camera – in order to make the problem tractable. In particular, we aim to predict the parts in an image sequence that are considered interesting by many viewers. To this end, we propose to combine several cues to assess interestingness in a novel manner that leads to increased performances when compared to averaged human ratings. Finally, we interpret and discuss the ability of computational techniques to spot these interesting events automatically. The relatively constrained setting allows for the discovery of crucial properties of interestingness and will pave the way for further, more challenging scenarios.

2. RELATED WORK

Psychological Perspective. In the mid-50s, Berlyne was among the first to seriously consider interest as psychologically relevant for human learning. In his seminal work [1] he introduced four collative variables, which affect interest: *novelty*, *uncertainty*, *conflict* and *complexity*. More recent research empirically validates and refines this theory, *e.g.* Chen *et al.* [5] declare novelty, challenge, attentional demand, exploitation intention and instant enjoyment as sources for interestingness (see [23] for a survey and comprehensive discussion). These theories have one limitation however, as they cannot explain why people respond differently. This is due to the fact that they implicitly trace interest to events rather than to interpretations and appraisals of events. Summarizing, these classical theories have been applied successfully in order to answer the questions what and why some things are interesting to almost everybody. As this is the main focus of the paper at hand, we only take into account these theories.

(Visual) Cognitive Perspective. The concept of interestingness has mostly been studied through understanding which visual stimuli can attract human attention [27]. This is often done by recording gaze patterns of humans watching images or videos (*e.g.*, [8, 10]). Despite a number of unsolved issues, there is common agreement that capturing human attention involves two fundamental properties of human information processing (*cf.* [22]): First, stimulus-based or bottom-up processing and secondly, memory-based or top-down processing. The main bottom-up factors that contribute to the (covert as well as overt) spatial allocation of visual attention are saliency of an object and novelty of an event. The seminal work of Treisman and Gelade [26] introduced the feature-integration theory which has been picked up frequently. For instance, Itti and Koch [12] included three stimulus features (orientation, intensity, and color), and received considerable agreement of their model with human gaze measurements. Motion and abrupt onset are the other features sometimes viewed as relevant in bottom-up processing. Despite the evidence pointing to the crucial role of bottom-up cues, it is obvious that they alone are insufficient to guide visual attention. In fact, it has been repeatedly shown that top-down processes can bias or even override bottom-up visual processing, *e.g.* [8, 25]. Top-down processing means that individuals are willfully able to track and search for relevant information, while ignoring irrelevant visual stimuli. This processing is strongly affected by task instruction, individual attentional resources, prior knowledge, and personal motivation or goals.

Summarizing, it has been shown that saliency and novelty clearly trigger visual attention. However, we argue that this is not necessarily equivalent to interestingness. If a person scans an image or a video in order to understand what is happening, this does not mean that she or he really considers the observations as interesting.

Computational Perspective. Different approaches have been proposed for the automatic detection of visual concepts that relate to interestingness. Applications include event detection (*e.g.*, [15]), video summarization (*e.g.*, [18]) or content-based image retrieval (*e.g.*, [6]). For example, Johnson and Hogg [15] refer to statistical *outliers* as "possible incidents of interest" or Stauffer and Grimson [24] claim many of their detections to be of "most interest". Other related terms that are often used include *surprise*, *saliency*, *abnormality* or *novelty*. The related techniques can be categorized as follows.

Abnormality Detection. In many abnormality detection algorithms, a model of normality is trained from frequent observations. Outliers to these models must be novel concepts and are identified as abnormal events. Typically, such approaches work well for fixed cameras, modeling the entire scene (*e.g.*, [31, 2, 13]) or the behavior of objects within this scene (*e.g.*, [15, 24]). In practice however, it is often unclear to what extent such anomalies are also perceived as interesting by human observers.

Attention Modeling. A simple example shows that "novel" and "interesting" are not always identical. In the white snow paradox, a TV-screen presenting a white noise signal is completely unpredictable and always remains novel, but is unattractive to viewers. This was already noticed in the beginnings of cognitive psychology [1, 30] and calls for a direct modeling of visual concepts which attract human attention. For example, Itti and Baldi came up with a theory of Bayesian surprise [11] or Schmidhuber and co-workers [21, 20] define interestingness as allowing for learning new things: "Neither the arbitrary nor the fully predictable is truly surprising or interesting – only data with still unknown but learnable statistical regularities are".

Interestingness as Category. Machine learning methods are successfully used for many vision tasks, such as object detection or recognition (*e.g.* detecting faces). Recent works widen these techniques to more complex and less well-defined tasks, for example *emotions* [17], *human memorability* [9] or *aesthetics* [6]. Furthermore, Weinshall *et al.* [29] introduced a novel concept based on disagreement of specific and general classifiers. In all these approaches, a machine learning method is provided with various, low level or specifically designed features. To this end, labeled training data must be available, which is in practice hard to gather in sufficient quality and quantity. For example, Dhar *et al.* [6]

use per-image "interestingness" scores from the photo sharing site *flickr*[2]. Yet, we have some doubts about the more general usefulness of these scores since they are based on properties including number of clicks and comments or popularity of the photographer, that lack independent human ratings.

3. HUMAN CONSENSUS BASE-LINE

The evaluation of theories and computational techniques always requires a reliable ground truth to compare against. In the case of "interestingness" this is clearly a highly subjective judgment and no truly correct answer can be expected from a single person. Therefore, we rely on the assessment by multiple people and establish a human consensus that serves as base-line for further investigations. The entire sequences as well as the obtained ratings are available on the authors' web-page.

3.1 Dataset

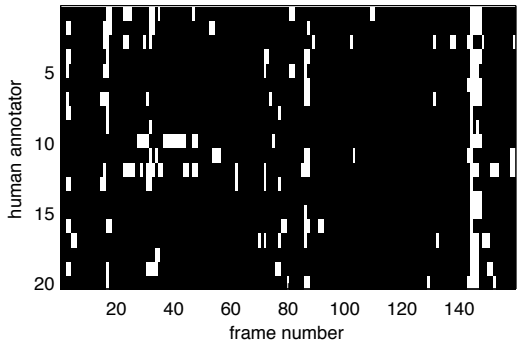
We chose 20 sequences from publicly available webcams, see Tab. 1. These image sequences were recorded over a long period of time, capturing various – possibly interesting – situations. The sequences present typical webcam and surveillance scenes, such as panoramas (*Seq. 1,4,11,17*), highways (*Seq. 5,18*), public squares (*Seq. 3,6,8,15*), urban scenes (*Seq. 10,14,20*), and some particular scenarios (boat rental, *Seq. 2*; stork nest, *Seq. 7*; beach, *Seq. 9*; construction site, *Seq. 12*; the Panama Canal, *Seq. 13*; a port, *Seq. 16*; and the Tower Bridge, *Seq. 19*). Image resolutions range from 352×288 (PAL) up to 420×315. Images were recorded during several days and usually sampled at one frame per hour. We manually selected representative sub-sequences, *e.g.* excluding very dark night images. Each image sequence consisted of 159 color images, continuously displayed at approximately 1 fps.

3.2 User Study

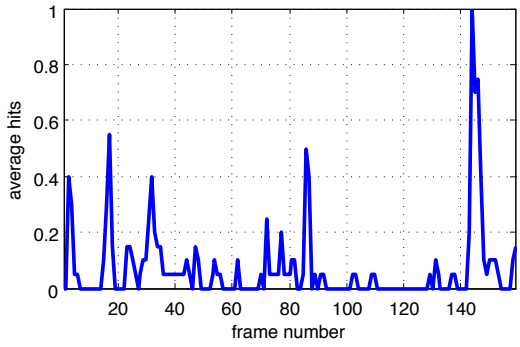
Setup. 26 male and 20 female test persons, aged between 18 and 47 and having normal or corrected vision participated in the test. They were instructed to watch the image sequences, press a button if they considered something interesting, and they had to press the button again to release the interestingness tag. No further instruction was given and hence the participants were free to judge what they considered as interesting. Furthermore, they were asked to rate the overall interestingness of every sequence at the end of the experiment, once they had seen all image sequences. In order to recall the sequences, the beginning of each sequence was replayed briefly, and ratings were asked in 7 levels (1 = very boring, 7 = very interesting). Test persons completed the task unwatched and they could stop when they reached their personal time budget. Sequences and their ordering were chosen randomly. Each participant evaluated between 6 and 10 sequences and all sequences were viewed by at least 20 persons.

Human Consensus. Let $a_t^{(i)}$ be the individual binary interestingness annotations of person i for image I_t for a particular sequence. If the user considers the frame as interesting $a_t^{(i)} = 1$, and 0 otherwise. The human consen-

[2]`http://www.flickr.com/explore/interesting/`, 2012/11/12.



(a) human annotations $a_t^{(i)}$



(b) human consensus s_t

(c) # 144 (d) # 17 (e) # 86

(f) # 32 (g) # 3 (h) # 72

(i) uninteresting (score = 0.00)

Figure 2: Example of user annotations (white = interesting) (a) and the obtained human consensus for *Seq. 1* (b). Interesting parts (c)-(h) include abnormal, unexpected events as well as "aesthetic" images. Images consistently considered as uninteresting (i) might have large variations at pixel level (*e.g.*, the sky region) but are semantically "normal".

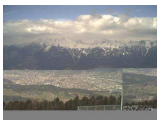
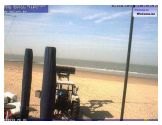
Seq.	1	2	3	4	5	6	7
typical image							
most interesting image							
s_{max}	1.00	0.75	0.50	0.75	0.45	0.95	0.70
$s(\mu \pm \sigma)$	0.07 ± 0.14	0.14 ± 0.15	0.10 ± 0.11	0.12 ± 0.14	0.06 ± 0.08	0.21 ± 0.25	0.19 ± 0.14
$\#s > 0.50$	4	5	0	3	0	26	7
$\#s < 0.25$	146	123	136	129	152	100	113
α_{st}	0.90	0.82	0.73	0.79	0.74	0.93	0.82
$r(\mu \pm \sigma)$	2.5 ± 1.2	3.6 ± 1.0	2.5 ± 1.1	3.4 ± 1.5	1.9 ± 0.8	4.0 ± 1.3	4.7 ± 1.7
Seq.	8	9	10	11	12	13	14
typical image							
most interesting image							
s_{max}	0.80	0.70	0.95	0.70	0.70	0.80	0.80
$s(\mu \pm \sigma)$	0.05 ± 0.11	0.07 ± 0.13	0.08 ± 0.17	0.13 ± 0.14	0.09 ± 0.13	0.16 ± 0.18	0.11 ± 0.17
$\#s > 0.50$	2	4	8	2	3	7	7
$\#s < 0.25$	152	143	139	125	144	114	135
α_{st}	0.87	0.88	0.92	0.75	0.80	0.85	0.88
$r(\mu \pm \sigma)$	2.7 ± 1.1	2.5 ± 1.1	2.8 ± 1.1	4.0 ± 1.3	2.2 ± 1.2	3.6 ± 1.1	2.8 ± 1.3
Seq.	15	16	17	18	19	20	average
typical image							
most interesting image							
s_{max}	0.65	0.80	0.65	0.55	0.70	0.80	0.74
$s(\mu \pm \sigma)$	0.13 ± 0.18	0.13 ± 0.15	0.12 ± 0.14	0.06 ± 0.10	0.11 ± 0.14	0.09 ± 0.13	0.11
$\#s > 0.50$	11	6	5	1	4	3	5
$\#s < 0.25$	126	124	134	147	135	140	133
α_{st}	0.86	0.82	0.78	0.78	0.83	0.82	0.83
$r(\mu \pm \sigma)$	2.9 ± 1.0	3.2 ± 1.4	2.9 ± 1.6	1.9 ± 0.7	3.5 ± 1.6	3.6 ± 1.6	3.0

Table 1: Dataset and statistics of the obtained human consensus base-line. For each sequence are shown: a typical (uninteresting) image, the most interesting image with its corresponding score, mean and variance of the scores, the number of interesting and uninteresting events, and an overall interestingness rating.

sus interestingness score is defined as the per-frame average of the individual annotations for a particular sequence, *i.e.*, $s_t = \frac{1}{N}\sum_i a_t^{(i)}$, were N is the number of participants who annotated this sequence. Similarly, the overall interestingness rating r of a sequence is the average of the individuals' overall ratings. If many individuals consider a frame interesting (*i.e.*, $s_t > 0.5$), we call this an interesting event. Hence, the most interesting events in a sequence can be ranked with respect to their human consensus interestingness score, that is the agreement among individuals.

Example Sequence. A detailed example is given in Fig. 2, showing the raw user annotations (a), the average

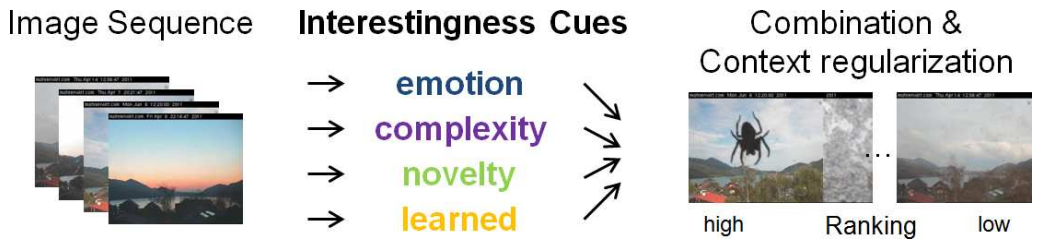


Figure 3: Given an image sequence, our aim is to automatically rank the images according to their interestingness. Psychological research suggests various cues that influence interestingness. In our approach, we use quite simple implementations for some of those cues and aggregate their outputs.

across persons – the established human consensus – (b) and examples of highly interesting (c)-(h) and uninteresting (i) frames. Interesting frames include surprising and unexpected events (the spider or snow on the lens) as well as aesthetic images (sunsets). Note that many other frames also might show large variations (especially in the sky region) but are consistently considered as uninteresting. Furthermore, at the very end of this sequence the camera settings were changed substantially (zoomed in). This change however, was mostly ignored by the viewers (Fig. 2 (i)). Hence, it seems that abstraction and semantic interpretation of what we expect is essential.

Dataset Overview. Tab. 1 summarizes results of the human responses. For each sequence are shown: a typical (uninteresting) image, the most interesting image with its corresponding human consensus interestingness score s_{max}, the number of interesting and uninteresting events, as well as the overall interestingness rating. Some events clearly appear interesting to many viewers, *e.g.* in *Seq. 6* many things are happening in the observed square (market, auto show, sports game, *etc.*). In each sequence however, there are frequent intervals that are consistently labeled as uninteresting.

3.3 Consistency

Per-frame Score. In order to quantify the consistency among the responses of test persons, we use standardized Cronbach's alpha. This widely used measure for the reliability of a psychometric test is defined as $\alpha_{st} = \frac{n\bar{r}}{1+(n-1)\bar{r}}$, where n is the number of persons and $\bar{r}$ the mean correlation between each of them.[3] As can be seen from Tab. 1, the measured consistency is generally high for most sequences (avg. $\bar{\alpha_{st}} = 0.83$, max. $\alpha_{st,max} = 0.93$). The responses of some sequences are less consistent (*e.g.*, *Seq. 11*, $\alpha_{st} = 0.75$). This seems due to the influence of individual preferences such as special cloud formations or sunsets.

Overall Rating. We use Sperman's rank coefficient ρ to assess the consistency across the participants overall ratings of the viewed sequences. This measure reflects how well two variables can be explained by a monotonic relation, therefore we first rank the annotated scenes according to the participants' ratings. Overall we achieved a mean rank coefficient of $\bar{\rho} = 0.30$ across all participants ($\rho_{max} = 0.94$; $\rho_{min} = -0.83$). This rating consistency is relatively weak, compared to the consistent annotations of per-frame interestingness. Remarkably, we spotted a more significant correlation $\bar{\rho}_{\mu_s} = 0.41$ between the average interestingness score μ_s of a sequence with its' ranking. Hence, having reliable interestingness scores per frame would also allow for a rough overall ranking.

Summarizing, the human consensus base-line of per-frame scores provides a solid basis and can be used to (i) build computational models and (ii) evaluate them.

[3]Literature suggests, $\alpha_{st} > 0.7$: acceptable; $\alpha_{st} > 0.8$: good; and $\alpha_{st} > 0.9$: excellent.

4. COMPUTATIONAL MODEL FOR INTERESTINGNESS

In this section we describe our approach for automatically predicting interestingness. More formally, given an image sequence $\mathbf{I} = [I_1, \ldots, I_n]$ we are looking for a ranking R of $\mathbf{I}$ with respect to its interestingness. The overview is sketched in Fig. 3 and consists of mainly two stages: (i) exploring various cues which serve as indicator for interestingness and (ii) combining these individual cues.

4.1 Cues for Interestingness

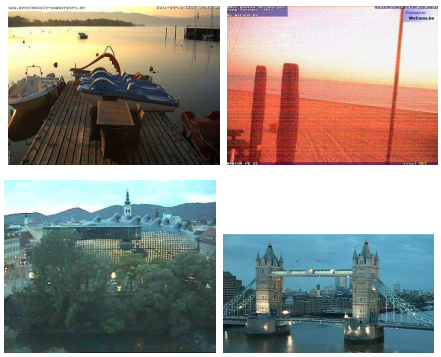
As reviewed in Sec. 2, psychological theories give us some ideas of what affects human interest. We select three cues, that are emotion, complexity and novelty. However, coming up with robust algorithms for these concepts is very challenging from a computer vision perspective. Due to the success of statistical and machine learning methods, we apply a learning technique as a fourth cue. Given an image, the learned model aims to directly predict its interestingness.

Features. Previous works have successfully used global scene descriptors as features to represent the images in tasks like scene characterization [13, 25], abnormality detection [2] or the rating of image memorability [9]. In the same vein, the features we use are raw pixel values, color histograms, histograms of oriented gradients (HOG), the Gist of the scene, and image self-similarity.

In the sequel we describe our implementations for the four cues. All these methods require specific parameter settings and critically depend on the employed distance measure and feature type. For the ease of clarity we restrict ourselves to report a few successful feature-algorithm combinations and parameter settings.

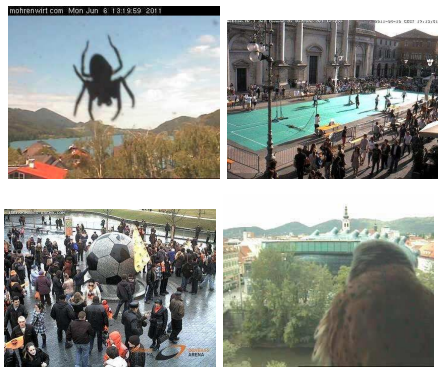
Each algorithm is supposed to return a score $\hat{s}_i \in \mathbb{R}$ for image I_i with respect to the interestingness, considering the entire image sequence $\mathbf{I}$, *i.e.*, $\hat{s}_i > \hat{s}_j$ iff I_i is considered to be more interesting than I_j. Examples of high ranked images for the different cues are shown in Fig. 4.

Emotion. Emotions can be characterized using the space of *pleasure*, *arousal* and *dominance* (see *e.g.* [16]).



(b) emotion

(d) complexity



(f) novelty

(h) learning

Figure 4: Typical examples of high ranked images for the individual cues obtained by our implementation.

We make use of empirical findings [28] in order to predict an emotion score from raw pixel values. As interesting events excite us (either positively or negatively), we propose to relate it to the arousal. Hence, the interestingness score is calculated as average over all pixels $p \in I_i$ as $\hat{s}_i^{(emo)} = \sum_p -0.31 \text{ brightness}(p) + 0.60 \text{ saturation}(p)$.

Complexity. To capture the complexity of an image, we quantify the visual structure present in the image. As a simple method, we use the the file-size of the image when encoded as PNG-image, *i.e.*, $\hat{s}_i^{(com)} = \text{bytes}(\text{encodePNG}(I_i))$.

Novelty. In the spirit of [31, 2], we perform outlier detection in the image stream. To this end, we use the Local Outlier Factor [3] with a 10-distance neighborhood. The predicted interestingness score $\hat{s}_i^{(nov)} = \text{LOF}_{10}(I_i, \mathbf{I})$ corresponds to the degree of being an outlier. As similarity measure we use sum of absolute differences $(\sum_p ||I_i(p) - I_j(p)||)^{-1}$ of the RGB-pixels values p within the two respective images i, j.

Learning. Finally we aim for directly learning a model of interestingness. Similar to [6, 9, 17] we train a classifier H to predict the interestingness score $\hat{s}_i^{(lea)} = H(I_i)$ for an image I_i. To this end, we extract Gist features and employ a Support Vector Regression (ν-SVR) [4][4] with an RBF kernel. The classifier is trained using the labeled data from the established human consensus baseline from all sequences except the sequence under test.

4.2 Combination

Model learning. The scores obtained from the respective cues are first normalized with respect to their mean and variance. Secondly, they are mapped into the interval $[0, 1]$ using a sigmoid function $\hat{s}_N = \frac{1}{1+exp(-a\hat{s}+b)}$. These normalized scores for emotion, complexity, novelty and learning, can be seen as high-level features. For combining them we train a simple linear model

$$\hat{s} = \mathbf{w}^\mathsf{T} \hat{\mathbf{s}}_N, \text{ with } \hat{\mathbf{s}}_N = [\hat{s}_N^{(emo)}, \hat{s}_N^{(com)}, \hat{s}_N^{(nov)}, \hat{s}_N^{(lea)}]^\mathsf{T} \quad (1)$$

The parameter a and b as well as the weights $\mathbf{w}$ were obtained by least square minimization[5] and cross-validation using the labeled data from the established human consensus baseline.

Regularization. So far, interestingness scores are obtained for each image in the image sequence independently. However, given the fact that we analyze an image sequence we can impose regularization terms which take this additional context into account.

We make use of the following two assumption: (i) visually similar images and (ii) temporally neighboring images should also have a similar interestingness score (*cf.* [14, 19], respectively). Please note that this is not automatically satisfied when using discriminatively trained classifiers, outlier detection methods and classifier combination methods as used in our case.

Graph based model. We cast the problem as transductive semi-supervised learning problem. By using a graph-based

[4]LIBSVM version 3.11, `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`, 2012/02/23

[5]We also trained a ν-SVR [4] on these high-level features, which lead to a similar performance.

approach [32] the individual images are represented as the nodes in the graph and the assumptions are encoded by the graph structure. More formally, let $\hat{\mathbf{s}} = [\hat{s}_1, \ldots, \hat{s}_n]$ be the vector of initial interestingness scores and let $\mathbf{W} = [w_{ij}]$ be the weight matrix encoding the graph structure, where w_{ij} are the weights connecting images i and j and are defined as follows:

(i) similar images (similar appearance). As weights w_{ij} we used the visual similarity of image i and j measured using sum of absolute differences on RGB pixel values, as described above. The weights are normalized by the maximum within the whole image sequence $\mathbf{I}$.

(ii) temporal consistency. In order to take the temporal consistency into account we added a constant c to the normalized weights $w_{i,i+1}, i = 1, \ldots, n-1$ of temporal neighboring images. In all our experiments we fixed $c = 0.25$ and obtained a good trade-off of temporal regularization (smoothing) and specificity (*i.e.*, avoiding over-smoothing).

The combined interestingness score $\hat{\mathbf{s}}$ is used as prior knowledge and the deduced labels $\bar{\mathbf{s}} = [\bar{s}_1, \ldots, \bar{s}_n]$ are inferred according to the graph structure as

$$\bar{\mathbf{s}} = (\mathbf{1} - (1-\eta)\mathbf{D}^{-1}\mathbf{W})^{-1}\eta\hat{\mathbf{s}}, \qquad (2)$$

where $\mathbf{D}$ is a diagonal matrix with $D_j = \sum_j w_{ij}$ and $\eta \in [0, 1]$ the regularization factor.

Ranking. The overall ranking R is determined by sorting the scores $\bar{\mathbf{s}}$ in descending order.

5. EXPERIMENTAL RESULTS

We first define the evaluation metrics and then present and discuss the results.

5.1 Evaluation of an Image Sequence

Top 3 Score (Top3). This metric quantifies how well the top 3 ranked frames of an image sequence analyzed by a computer vision technique capture the human interestingness score. To this end, we define the fraction $Top3 := \frac{\sum_{k \in A^{Top3}} s_k}{\sum_{k \in S^{Top3}} s_k}$, where s_k is the human consensus score of frame k, A^{Top3} the set of the top 3 ranked samples determined by the algorithm, and S^{Top3} the set of top 3 ranked samples of the human consensus, respectively. $Top3 \in [0, 1]$, where a higher value corresponds to a better performance of the algorithm.

In the same manner, other scores, *e.g. Top10* or *Bottom3*, can be defined. Yet, due to the low number of interesting events, we experienced a good characterization with *Top3*. Furthermore, *bottom* scores as well as global rankings (*e.g.*Spermans ρ) are inappropriate measures as the dataset is highly unbalanced (cf. Tab. 1).

Average Precision (AP). As we are mainly interested in robustly spotting interesting events, we also perform a recall-precision analysis, using the average precision value AP [7]. Interesting events are defined to have a consensus score above 0.5. On the other hand, non-interesting frames are those for which the score is below 0.25 (*cf.* Tab. 1). Images with a score in between are ignored for the evaluation, due to the fact that also humans do not agree well on them.

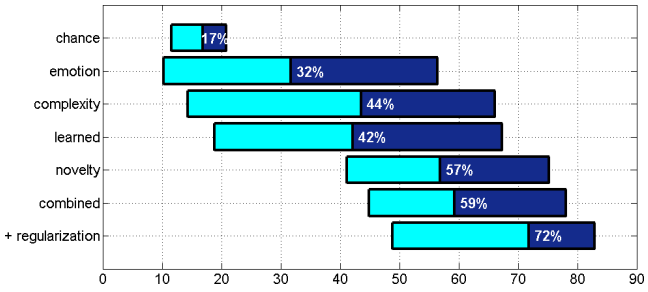


Figure 5: Boxplot of *Top3* performance (median values are printed) for the individual cues used and the proposed combinations with respect to the human consensus baseline. All individual cues are clearly above chance level, headed by novelty. However, not all novel images are necessarily interesting nor are all interesting images novel, thereby calling for a combination of these cues. This combination shows improved performance, especially when applying regularization.

5.2 Results and Discussions

Results per-sequence and overall statistics are presented in Tab. 2 and shown in Fig. 5.

Individual cues. The results for all individual cues are clearly above chance level, proving their usefulness for the task. The most dominant cue, however, is novelty. With respect to novelty detection, using low resolution (32×32 pixels) images already shows a good performance. We tried various other image resolutions, but sizes above 32×32 pixels yield marginal effects, whereas lower resolutions resulted in significantly decreased performance. Similar results were observed when using other features or other outlier detection methods[6]. On the one hand, this observations validates to some extent the assumptions used in many abnormality detection systems. On the other hand, the dominance of novelty might be also due to the current understanding and implementation of outlier detection methods. For the other cues unfortunately no satisfying implementations exists, which might be due to overly simple visual features and/or the limited amount of training data.

Combination. In many cases abnormal events are considered interesting. This is true for some large (unexpected) changes in the image sequence, as for instance the spider in *Seq. 1* or the crowd appearing in *Seq. 8*. On the other hand, abnormality detection can mislead the interestingness output, in particular in cases that are easily interpretable for humans. This occurs for large shadows, different weather conditions, camera failures or slightly shifted/zoomed camera views. Such cases call for the inclusion of other cues. Furhtermore, some very subtle but semantically meaningful changes make an image interesting, as for instance the appearing egg in *Seq. 7* or the crane in *Seq. 10*. In this respect, also the entire scene plays a role, as the two extreme

[6]We also experimented with *DB(p,D)*-Outliers, one-class Support Vector Machine and Meaningful Nearest Neighbors. Gaussian Mixture Models and Principle Component Analysis operating on pixel level do not perform that well, mainly due to the large data variance in the time-lapse image sequences.

cases of a very boring motorway in *Seq. 18* and the highly entertaining stork nest in *Seq. 7* show. Effectively handling such cases calls for a more semantic interpretation to filter out irrelevant information while sharpening the focus on relevant regions or objects. In the extreme case a complete understanding of the entire visual scene with all its objects and interactions might be necessary. However, our results show that the use of simple combination of even basic cues already yield affective results on average, for both *Top3* and *AP* measures (*Top3* med.: 0.72, *AP* med.: 0.36). In particular, integrating visual and temporal consistency turned out to have a significant effect. Some visual examples of success and failure cases are depicted in Fig. 6.

Relation to other works. To show the benefit of our method, we compare it to other related works. First, we have implemented a technique based on [20], where the interestingness score $\hat{s}_t = -\frac{\partial^2}{\partial t^2} l_C([I_1, \ldots, I_t])|_t$ is defined as the negative second derivative of an on-line compression capacity (l_C is the length of the resulting string when compressing the history of all images upon time t with a fixed compressor C). We therefore concatenate images one by one and use the file size of the resulting PNG image as compression measure. This technique exhibits inferior performance (*Top3* med.: 0.36, *AP* med.: 0.10). Our interpretation is that outliers and abnormal events are in many cases perceived as interesting by human observers, but such abnormalities are explicitly excluded here. Second, we also did not achieve good results using the theory of Bayesian surprise [11] (*Top3* med.: 0.17, *AP* med.: 0.07, which is around chance level). It appears that the goal there is a different one, *i.e.* detecting the most surprising regions in high frame-rate videos, whereas our goal is to label interesting frames in the context of (low frame-rate) image sequences.

6. CONCLUSION

We assessed the capacity of computational approaches to capture "interestingness" in image sequences. Therefore, we recorded moments of human interest and aggregated them to a consensus, which is used for training and evaluating our approach. Motivated by psychological findings, we proposed to use four different cues, and showed an effective way to combine them. The image sequences specifies a context in which humans judge events as interesting and therefore outlier detection techniques permit to capture a substantial fraction of interestingness already. While many abnormal events are consistently considered to be interesting, also a large portion of them are not, such as camera failures or different cloud formations. On the other hand, statistically well-explained, normal events might still be interesting, *e.g.* raising of the Tower Bridge. This calls for a combination of abnormality detection and methods relying on semantic image interpretation. We have demonstrated that – while still remaining far from a semantic interpretation – a number of basic cues and their straight forward combination already let us for a long way in mimicking human interest responses.

This said, this work is only a first attempt to quantify visual interestingness. In future work it will be important to add semantic cues and to exploit promissing additional cues (*e.g.*examine the relationship between visual interestingness and other measures, such as memorability or image quality), widen the scope to more general settings, and taking the specific preferences of a particular observer into account.

7. ACKNOWLEDGMENTS

This research was supported by the Swiss CTI under project no. 15769.1, "Relevance feedback mechanisms for video surveillance".

8. REFERENCES

[1] D. Berlyne. *Conflict, arousal, and curiosity*. McGraw-Hill, 1960.

[2] M. Breitenstein, H. Grabner, and L. Van Gool. Hunting nessie: Real time abnormality detection from webcams. In *Proc. IEEE WS on Visual Surveillance*, 2009.

[3] M. Breunig, H. Kriegel, R. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2000.

[4] C. Chang and C. Lin. LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27):1–27, 2011.

[5] A. Chen, P. Darst, and R. Pangrazi. An examination of situational interest and its sources. *British Journal of Educational Psychology*, 71:383–400, 2001.

[6] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proc. CVPR*, 2011.

[7] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010.

[8] J. Henderson. Human gaze control during real-world scene perception. *TRENDS in Cognitive Science*, 7(11):498–504, 2003.

[9] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Proc. CVPR*, 2011.

[10] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *Proc. CVPR*, 2005.

[11] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49:1295–1306, 2009.

[12] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[13] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. In *Proc. CVPR*, 2007.

[14] V. Jain and E. Learned-Miller. Online domain adaption of a pre-trained cascade of classifiers. In *Proc. CVPR*, 2011.

[15] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. In *Proc. BMVC*, 1995.

[16] P. Lang, M. Bradley, and B. Cuthbert. International affective picture system (iaps): Affective ratings of pictures and instruction manual. Technical report, Technical Report A-8. University of Florida, 2008.

[17] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proc. ACM Multimedia*, 2010.

[18] A. Money and H. Agiu. Video summarisation: A conceptual framework and survey of the state of the art. *J. Visual Communication and Image Representation*, 19:121–143, 2008.

[19] F. Nater, H. Grabner, and L. Van Gool. Temporal relations in videos for unsupervised activity analysis. In *Proc. BMVC*, 2011.

[20] T. Schaul, L. Pape, T. Glasmachers, V. Graziano, and J. Schmidhuber. Coherence progress: A measure of interestingness based on fixed compressors. In *In Proc. Conf. on Artificial General Intelligence*, 2011.

[21] J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. Technical report, TU-Munich, 2009.

sequence	chance	emotion	complexity	learned	novelty	combined	+ regularization
1	*0.09*	0.27	0.14	0.71	**0.88**	**0.88**	**0.88**
	0.04	0.10	0.15	0.64	0.83	**0.88**	0.84
2	*0.19*	0.10	0.10	0.03	0.41	**0.59**	**0.59**
	0.08	0.18	0.05	0.07	0.31	0.34	**0.36**
3	*0.23*	0.07	0.71	**1.00**	0.71	0.86	0.86
	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	*0.20*	**0.56**	0.10	0.46	0.18	0.13	0.26
	0.03	**0.18**	0.02	0.15	0.10	0.10	0.11
5	*0.14*	0.13	0.13	0.04	0.09	0.09	0.04
	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	*0.21*	0.36	0.68	0.68	0.79	**0.81**	0.74
	0.22	0.54	0.63	**0.80**	0.41	0.67	0.68
7	*0.28*	0.50	0.12	0.31	0.62	0.48	**0.76**
	0.12	0.25	0.09	0.05	0.26	0.23	**0.28**
8	*0.09*	**1.00**	**1.00**	0.21	0.74	0.62	**1.00**
	0.02	**1.00**	**1.00**	0.11	**1.00**	0.67	**1.00**
9	*0.13*	0.10	0.44	0.67	0.64	0.46	**0.72**
	0.04	0.04	0.11	0.15	**0.29**	0.27	0.15
10	*0.11*	0.06	0.45	0.06	0.60	0.60	**0.79**
	0.09	0.10	0.18	0.08	0.34	0.35	**0.36**
11	*0.21*	**0.66**	0.29	0.17	0.37	0.37	0.40
	0.06	0.04	0.02	0.03	**0.10**	0.05	0.04
12	*0.14*	**0.74**	0.44	0.36	0.54	0.38	0.72
	0.05	**0.68**	0.32	0.34	0.43	0.45	0.52
13	*0.20*	0.55	0.64	**0.71**	0.29	0.55	0.55
	0.11	0.28	**0.41**	0.35	0.14	0.25	0.17
14	*0.13*	0.57	0.43	0.61	**0.96**	**0.96**	0.87
	0.10	0.29	0.05	0.18	**0.90**	0.83	0.85
15	*0.23*	0.08	**0.92**	0.90	0.44	0.44	0.44
	0.12	0.14	**0.67**	0.57	0.10	0.22	0.22
16	*0.18*	0.10	0.14	0.50	0.76	0.76	**0.90**
	0.07	0.06	0.14	0.38	**0.93**	0.86	0.85
17	*0.19*	0.16	0.49	**0.62**	0.54	0.54	0.54
	0.07	0.05	0.24	**0.49**	0.31	0.39	0.36
18	*0.14*	0.28	**0.69**	0.38	0.41	**0.69**	0.41
	0.05	0.25	**0.50**	0.06	0.33	0.33	0.33
19	*0.16*	0.70	0.35	0.15	**0.80**	**0.80**	**0.80**
	0.11	0.28	**0.29**	0.03	0.28	0.28	0.27
20	*0.13*	0.40	0.40	0.38	0.48	**0.69**	**0.69**
	0.03	0.38	0.18	0.37	0.38	**0.59**	0.50
median	*0.17*	0.32	0.44	0.42	0.57	0.59	**0.72**
	0.07	0.22	0.18	0.16	0.32	0.35	**0.36**

Table 2: Results for automatically determining interestingness and spotting interesting events in image sequences (see also Fig 5). For each sequence, the first row shows the *Top3* score and the second row shows the *AP*. Best results are printed in bold face.

[22] W. Schneider and R. Shiffrin. Controlled and automatic human information processing: Detection, search, and attention. *Psychological Review*, 84:1–66, 1977.

[23] P. Silvia. *Exploring the psychology of interest*. Oxford University Press, 2006.

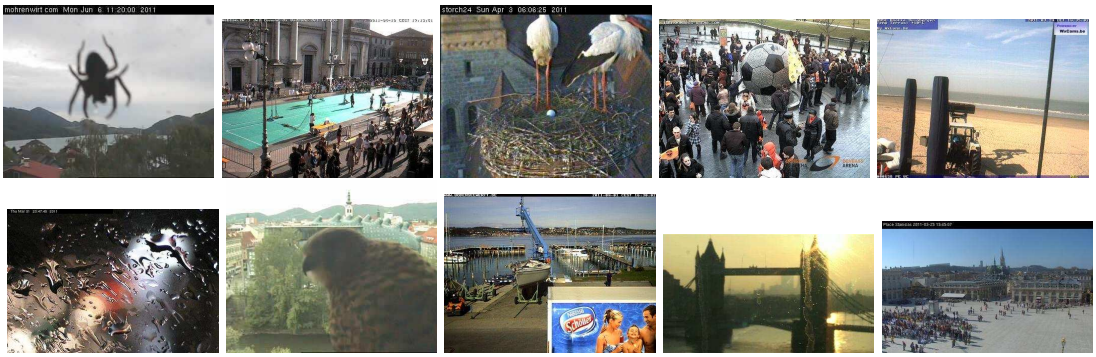
[24] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22(8):747–757, 2000.

[25] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113(4):766–786, 2006.

[26] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[27] J. Tsotsos, L. Itti, and G. Rees. A brief and selective history of attention. In *Neurobiology of Attention*. Elsevier, 2005.

[28] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394–409, 1994.



(a) **true positives** ($s_t > 0.5$, I_t within the top 5 of our ranking)

(b) **true negatives** ($s_t < 0.25$, I_t within the bottom half of our ranking)

(c) **false positives** ($s_t < 0.25$, I_t within the top 5 of our ranking)

(d) **false negatives** ($s_t > 0.5$, I_t within the bottom half of our ranking)

Figure 6: Typical examples of correctly (a,b) and incorrectly (c,d) classified images. False positives are often caused by dominant weather conditions or shadows, whereas false negatives mostly depict some very subtle and highly semantic concepts, such as the pigeons in the storks nest or raising of the Tower Bridge.

[29] D. Weinshall, A. Zweig, H. Hermansky, S. Kombrink, F. Ohl, J. Anemüller, J. Bach, L. Van Gool, F. Nater, T. Pajdla, M. Havlena, and M. Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. *PAMI*, 2012.

[30] W. Wundt. *Grundzüge der physiologischen Psychologie*. Engelmann, Leipzig, 1874.

[31] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. CVPR*, 2004.

[32] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. ICML*, 2003.